BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e35902
doi: 10.3897/biss.3.35902

OPEN ACCESS

Conference Abstract

The Pensoft Data Publishing Workflow: The FAIRway from articles to Linked Open Data

Lyubomir Penev[‡], Teodor Georgiev[§], Viktor Senderov[§], Mariya Dimitrova[§], Pavel Stoev[|]

‡ Pensoft Publishers & Bulgarian Academy of Sciences, Sofia, Bulgaria
§ Pensoft Publishers, Sofia, Bulgaria
| National Museum of Natural History and Pensoft Publishers, Sofia, Bulgaria

Corresponding author: Lyubomir Penev (penev@pensoft.net)

Received: 03 May 2019 | Published: 13 Jun 2019

Citation: Penev L, Georgiev T, Senderov V, Dimitrova M, Stoev P (2019) The Pensoft Data Publishing Workflow: The FAIRway from articles to Linked Open Data. Biodiversity Information Science and Standards 3: e35902. https://doi.org/10.3897/biss.3.35902

Abstract

As one of the first advocates of open access and open data in the field of biodiversity publishiing, Pensoft has adopted a multiple data publishing model, resulting in the ARPHA-BioDiv toolbox (Penev et al. 2017). ARPHA-BioDiv consists of several data publishing workflows and tools described in the Strategies and Guidelines for Publishing of Biodiversity Data and elsewhere:

1. Data underlying research results are deposited in an external repository and/or published as supplementary file(s) to the article and then linked/cited in the article text; supplementary files are published under their own DOIs and bear their own citation details.
2. Data deposited in trusted repositories and/or supplementary files and described in data papers; data papers may be submitted in text format or converted into manuscripts from Ecological Metadata Language (EML) metadata.
3. Integrated narrative and data publishing realised by the Biodiversity Data Journal, where structured data are imported into the article text from tables or via web services and downloaded/distributed from the published article.
4. Data published in structured, semanticaly enriched, full-text XMLs, so that several data elements can thereafter easily be harvested by machines.

© Penev L et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

5. Linked Open Data (LOD) extracted from literature, converted into interoperable RDF triples in accordance with the OpenBiodiv-O ontology (Senderov et al. 2018) and stored in the OpenBiodiv Biodiversity Knowledge Graph.

The above mentioned approaches are supported by a whole ecosystem of additional workflows and tools, for example: (1) pre-publication data auditing, involving both human and machine data quality checks (workflow 2); (2) web-service integration with data repositories and data centres, such as Global Biodiversity Information Facility (GBIF), Barcode of Life Data Systems (BOLD), Integrated Digitized Biocollections (iDigBio), Data Observation Network for Earth (DataONE), Long Term Ecological Research (LTER), PlutoF , Dryad, and others (workflows 1,2); (3) semantic markup of the article texts in the TaxPub format facilitating further extraction, distribution and re-use of sub-article elements and data (workflows 3,4); (4) server-to-server import of specimen data from GBIF, BOLD, iDigBio and PlutoR into manuscript text (workflow 3); (5) automated conversion of EML metadata into data paper manuscripts (workflow 2); (6) export of Darwin Core Archive and automated deposition in GBIF (workflow 3); (7) submission of individual images and supplementary data under own DOIs to the Biodiversity Literature Repository, BLR (workflows 1-3); (8) conversion of key data elements from TaxPub articles and taxonomic treatments extracted by Plazi into RDF handled by OpenBiodiv (workflow 5).

These approaches represent different aspects of the prospective scholarly publishing of biodiversity data, which in a combination with text and data mining (TDM) technologies for legacy literature (PDF) developed by Plazi, lay the ground of an entire data publishing ecosystem for biodiversity, supplying FAIR (Findable, Accessible, Interoperable and Reusable data to several interoperable overarching infrastructures, such as GBIF, BLR, Plazi TreatmentBank, OpenBiodiv and various end users.

Presenting author

Lyubomir Penev

References

- Penev L, Georgiev T, Geshev P, Demirov S, Senderov V, Kuzmova I, Kostadinova I, Peneva S, Stoev P (2017) ARPHA-BioDiv: A toolbox for scholarly publication and dissemination of biodiversity data based on the ARPHA Publishing Platform. Research Ideas and Outcomes 3: e13088. https://doi.org/10.3897/rio.3.e13088
- Senderov V, Simov K, Franz N, Stoev P, Catapano T, Agosti D, Sautter G, Morris RA, Penev L (2018) OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system. Journal of Biomedical Semantics 9 (1): 5. https://doi.org/10.1186/s13326-017-0174-5